HERACLES: Hyperbolic Embeddings for Reconstruction and Analysis of Lineages

Gilad Turok

Columbia University, Pe'er Lab
Supervised by Sitara Persad

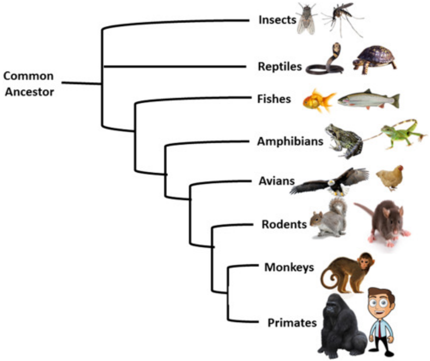
Semester Report
November 29, 2022

Outline

1. Background
2. Related Work
3. HERACLES Method
4. Conclusion

Background: Phylogeny

- **Phylogeny:** evolutionary development and diversification
- **Phylogenetic Tree:**

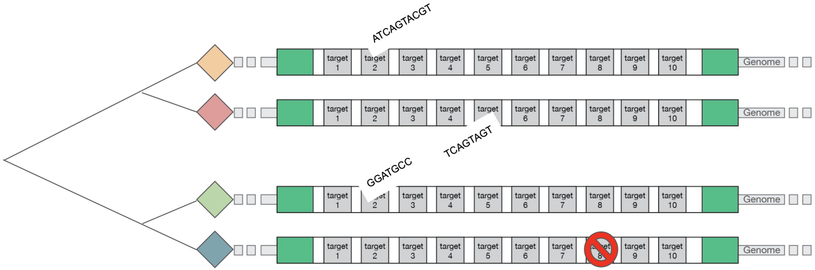


- **Idea:** Use inherited genetic alterations to infer phylogenetic trees

Background: CRISPR/Cas9

- Use CRISPR-Cas9 to precisely insert or delete genes in a cell at specific *target sites*
- These alterations will propogate through cell as they evolve
- In ”leaf cells”, examine the genes at target sites to reconstruct the cell lineage

Background: Character Matrix



Cells	Targets									
🔶	0	1	0	0	0	0	0	0	0	0
🟥	0	0	0	0	1	0	0	0	0	0
🟩	0	2	0	0	0	0	0	0	0	0
🔷	0	0	0	0	0	0	0	-	0	0

Figure from Anthony's presentation

Related Work: Methods for Lineage Tracing

Distance-Based Methods:

- Define pairwise distance function for two cells $d(c_1, c_2)$ that captures how ”far away” their sequences are
- Learn all-to-all distance matrix based D on d
- Construct tree from distance matrix D

Related Work: Methods for Lineage Tracing

Distance-Based Methods:

- Define pairwise distance function for two cells $d(c_1, c_2)$ that captures how ”far away” their sequences are
- Learn all-to-all distance matrix based D on d
- Construct tree from distance matrix D

Maximum Likelihood:

- Assume statistical model of DNA sequence evolution
- Infer probability distribution for diffirerent configigurations of phylogenetic trees
- Optimize over continous branch lengths and *discrete* tree topologies

Related Work: Methods for Lineage Tracing

Distance-Based Methods:

- Define pairwise distance function for two cells $d(c_1, c_2)$ that captures how ”far away” their sequences are
- Learn all-to-all distance matrix based D on d
- Construct tree from distance matrix D

Maximum Likelihood:

- Assume statistical model of DNA sequence evolution
- Infer probability distribution for diffirerent configigurations of phylogenetic trees
- Optimize over continous branch lengths and *discrete* tree topologies

Maximum Parsimony: identify phylogenetic tree with smallest number of evolutionary events to explain sequence data
Bayesian Inference: similar to maximum likelihood, but with Bayesian priors

Related Work: Wilson Paper

- **Paper:** *Learning Phylogenetic Trees as Hyperbolic Point Configurations* by Benjamin Wilson, 2021
- **Idea:** *Distance-based methods* in hyperbolic space can approximate *maximum likelihood methods*
- **Motivation:** Can embed a tree into hyperbolic space with an error bounded by 2δ

Theorem (Relaxed 4 Point Condition)

Let $\delta \geq 0$. A metric space (X, d) is said to be δ-hyperbolic if, for all $w, x, y, z \in X$,

$$d(x,w) + d(y,z) \leq \max\left\{d(x,y) + d(z,w),\ d(x,z) + d(y,w)\right\} + 2\delta$$

Related Work: Wilson Paper

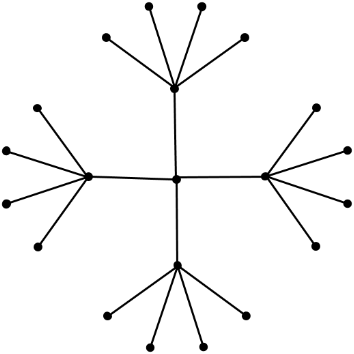


Figure: Euclidean Tree-Embedding

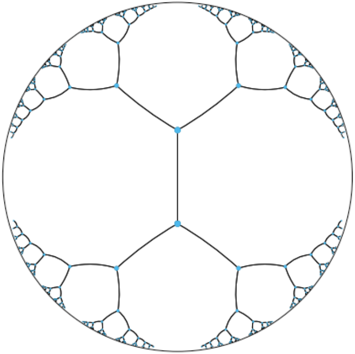


Figure: Hyperbolic Tree-Embedding

Related Work: Wilson Paper

For a Riemannian manifold $M = \mathbb{R}^m$, let a point configuration $\mathbf{x} = x_1 \dots x_n \in M$ with distance function d

For N cells and sequence data Θ of length L, let $\mathcal{L}_{ij}(t \mid \Theta)$ represent the likelihood of an evolutionary distance of t between cells i, j given sequence data Θ

Related Work: Wilson Paper

For a Riemannian manifold $M = \mathbb{R}^m$, let a point configuration $\mathbf{x} = x_1 \dots x_n \in M$ with distance function d

For N cells and sequence data Θ of length L, let $\mathcal{L}_{ij}(t \mid \Theta)$ represent the likelihood of an evolutionary distance of t between cells i, j given sequence data Θ

Our objective *logalike* function is

$$\mathbf{l}(\mathbf{x}) = \frac{1}{L} \sum_{i \neq j} \log \mathcal{L}_{ij}(d(x^i, x^j) \mid \Theta)$$

Related Work: Wilson Paper

For a Riemannian manifold $M = \mathbb{R}^m$, let a point configuration $\mathbf{x} = x_1 \ldots x_n \in M$ with distance function d

For N cells and sequence data Θ of length L, let $\mathcal{L}_{ij}(t \mid \Theta)$ represent the likelihood of an evolutionary distance of t between cells i, j given sequence data Θ

Our objective *logalike* function is

$$\mathbf{l}(\mathbf{x}) = \frac{1}{L} \sum_{i \neq j} \log \mathcal{L}_{ij}(d(x^i, x^j) \mid \Theta)$$

Unlike typical log-likelihood on treespace, $\mathbf{l}(\mathbf{x})$ is on a Riemannian manifold and is differentiable in $\mathbf{x}$

Can compute $\nabla_{x^i} \mathbf{l}(x^i)$ to learn ideal point configuration for infering phylogenetic tree

Related Work: Wilson Paper

Let σ_i be the base/state at target site σ for cell i

Assume evolution is a continous time markov chain with transition probablities P, infinitesmial generator Q, and stationary distribution π.

Specifically let $P_{\sigma_i \sigma_j}(t)$ represent the probability of observing state σ_j t time steps after observing state σ_i under the Jukes-Cantor model of mutation:

$$P_{ab} = \begin{cases} \frac{1}{4} + \frac{3}{4} e^{-4t/3} := P_{\text{diag}} & \text{if } a = b \\ \frac{1}{4} - \frac{1}{4} e^{-4t/3} := P_{\backslash\text{diag}} & \text{otherwise} \end{cases}$$

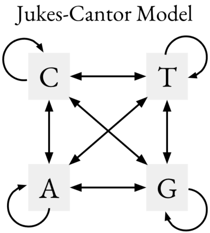


Figure from Sitara

Related Work: Wilson Paper

Objective logalike function:

$$\mathbf{l}(\mathbf{x}) = \frac{1}{L}\sum_{i\neq j} \log \mathcal{L}_{ij}(d(x^i, x^j) \mid \Theta)$$

Likelihood under assumed model of mutation

$$\mathcal{L}_{ij}(t) = \prod_{\text{sites } \sigma} \pi_{\sigma_i} P_{\sigma_i\sigma_j}(t)$$

$$\log \mathcal{L}_{ij}(t) = \sum_{\text{sites } \sigma} \log P_{\sigma_i\sigma_j}(t) + C$$

Expression for gradient with respect to cell *i*

$$\nabla_{x^i}\mathbf{l}(x^i) = \frac{1}{L}\sum_{\text{cells } j}\sum_{\text{sites } \sigma} \frac{(QP(d_{ij}))_{\sigma_i,\sigma_j}}{P(d_{ij})_{\sigma_i\sigma_j}} \nabla_{x^i} d_{x^j}$$

HERACLES Method: Model of Mutation

Accumulation of CRISPR-Cas9 mutations evolves as a continous time Markov Chain

Use a different model of mutation because known ancestral state, mutations cannot be undone, etc

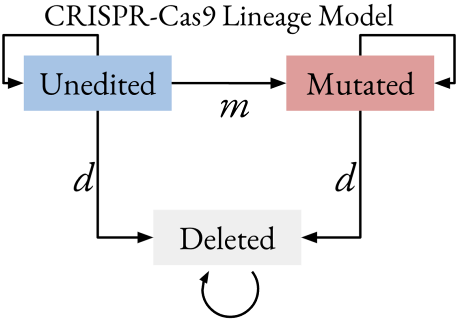


Figure from Sitara

HERACLES Method

For a cell c at site σ, possible mutation states are unedited ϕ_c, mutations $1 \ldots M_\sigma$, or deleted D:

$$\sigma_c \in \{\phi_c, 1, 2 \ldots M_\sigma, D\}$$

For mutation states a, b, let $P^\sigma_{ab}(t)$ be the conditional probability of observing state b after t time units have passed since observing state a.

HERACLES Method

Because of time-irrervsersibility, distance betwen cells i, j at site σ is determined by distance from common ancestor to each cell.

Let $A(\sigma_i, \sigma_j)$ be the set of all possible ancestors of cells i, j at site σ. Let π_a be the prior of observing ancestral state $a \in A$

$$\mathcal{L}_{ij}(t) = \prod_{\text{sites } \sigma} \sum_{a \in A(\sigma_i, \sigma_j)} \pi_a \, P_{a\sigma_i}(\frac{t}{2}) \, P_{a\sigma_j}(\frac{t}{2})$$

Biological constraints ensure set of ancestral states is small

HERACLES Method

The infinitesmial generator Q is defined below

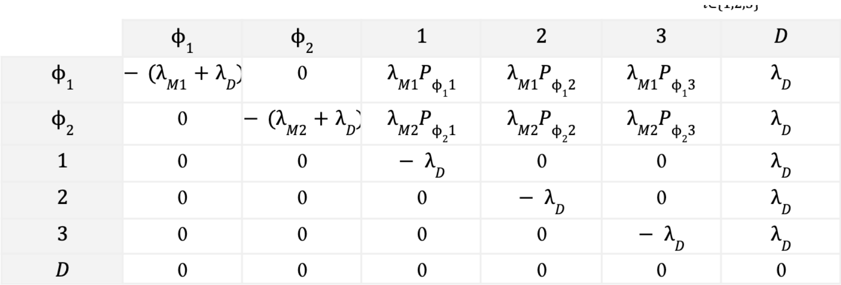
The time to leave the unedited state and transition to a mutated or deleted state follows an exponential distribution with rate $\gamma = \gamma_M + \gamma_D$

Given that a transition occurs to another character state, the transition to the observed character state C is $p_{\phi_\sigma, c}$

After transitioning to a character state, it may transition to a deleted state with rate λ_D

HERACLES Method

Infentiesmal generator Q is defined below with $\sum_i P_{\phi i} = 1$



$i\in\{1,2,3\}$

	ϕ_1	ϕ_2	1	2	3	D
ϕ_1	$-(\lambda_{M1} + \lambda_D)$	0	$\lambda_{M1}P_{\phi_1 1}$	$\lambda_{M1}P_{\phi_1 2}$	$\lambda_{M1}P_{\phi_1 3}$	λ_D
ϕ_2	0	$-(\lambda_{M2} + \lambda_D)$	$\lambda_{M2}P_{\phi_2 1}$	$\lambda_{M2}P_{\phi_2 2}$	$\lambda_{M2}P_{\phi_2 3}$	λ_D
1	0	0	$-\lambda_D$	0	0	λ_D
2	0	0	0	$-\lambda_D$	0	λ_D
3	0	0	0	0	$-\lambda_D$	λ_D
D	0	0	0	0	0	0

Probability transition matrix P is derived from Q:

$$P(t) = \mathrm{expm}(Q * t)$$

HERACLES Method

Objective logalike function:

$$\mathbf{l}(\mathbf{x}) = \frac{1}{L} \sum_{i \neq j} \log \mathcal{L}_{ij}(d(x^i, x^j) \mid \Theta)$$

Likelihood under assumed model of mutation

$$\mathcal{L}_{ij}(t) = \prod_{\text{sites } \sigma} \sum_{a \in A(\sigma_i, \sigma_j)} \pi_a \, P_{a\sigma_i}(\frac{t}{2}) \, P_{a\sigma_j}(\frac{t}{2})$$

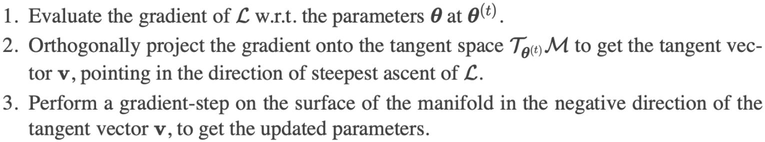
$$\log \mathcal{L}_{ij}(t) = \sum_{\text{sites } \sigma} \log \sum_{a \in A(\sigma_i, \sigma_j)} \pi_a \, P_{a\sigma_i}(\frac{t}{2}) \, P_{a\sigma_j}(\frac{t}{2})$$

Expression for gradient with respect to cell i

$$\nabla_{x^i} \mathbf{l}(x^i) = \frac{1}{L} \sum_{\text{cells } j} \sum_{\text{sites } \sigma} \frac{(QP(d_{ij}))_{\sigma_i, \sigma_j}}{P(d_{ij})_{\sigma_i \sigma_j}} \nabla_{x^i} d_{x^j}$$

HERACLES Method

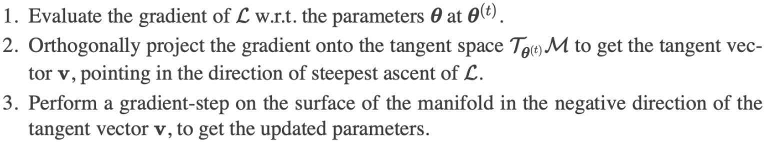
Focused on implementing Riemannian SGD



1. Evaluate the gradient of $\mathcal{L}$ w.r.t. the parameters $\boldsymbol{\theta}$ at $\boldsymbol{\theta}^{(t)}$.
2. Orthogonally project the gradient onto the tangent space $\mathcal{T}_{\boldsymbol{\theta}^{(t)}}\mathcal{M}$ to get the tangent vector $\mathbf{v}$, pointing in the direction of steepest ascent of $\mathcal{L}$.
3. Perform a gradient-step on the surface of the manifold in the negative direction of the tangent vector $\mathbf{v}$, to get the updated parameters.

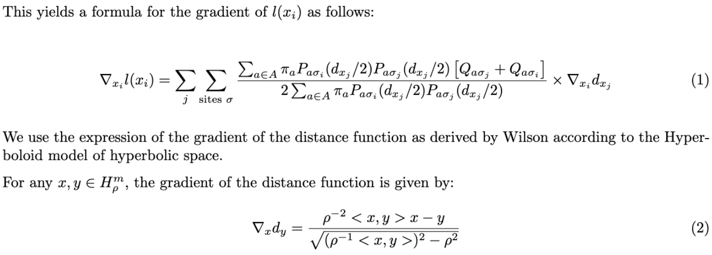
HERACLES Method

Focused on implementing Riemannian SGD



1. Evaluate the gradient of $\mathcal{L}$ w.r.t. the parameters $\boldsymbol{\theta}$ at $\boldsymbol{\theta}^{(t)}$.
2. Orthogonally project the gradient onto the tangent space $\mathcal{T}_{\boldsymbol{\theta}^{(t)}}\mathcal{M}$ to get the tangent vector $\mathbf{v}$, pointing in the direction of steepest ascent of $\mathcal{L}$.
3. Perform a gradient-step on the surface of the manifold in the negative direction of the tangent vector $\mathbf{v}$, to get the updated parameters.

Already have gradient in closed form



This yields a formula for the gradient of $l(x_i)$ as follows:

$$\nabla_{x_i} l(x_i) = \sum_{j} \sum_{\text{sites } \sigma} \frac{\sum_{a \in A} \pi_a P_{a\sigma_i}(d_{x_j}/2) P_{a\sigma_j}(d_{x_j}/2) \left[Q_{a\sigma_j} + Q_{a\sigma_i}\right]}{2 \sum_{a \in A} \pi_a P_{a\sigma_i}(d_{x_j}/2) P_{a\sigma_j}(d_{x_j}/2)} \times \nabla_{x_i} d_{x_j} \tag{1}$$

We use the expression of the gradient of the distance function as derived by Wilson according to the Hyperboloid model of hyperbolic space.

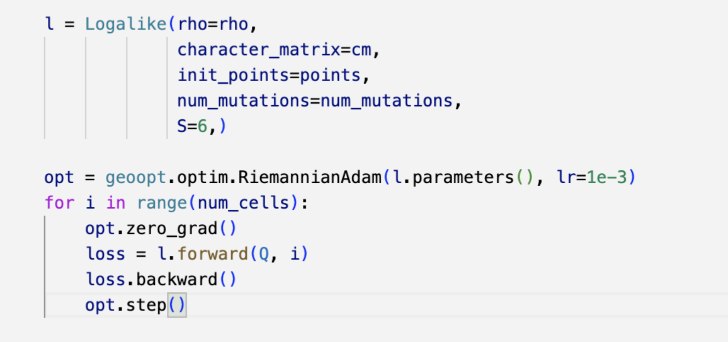
For any $x, y \in H^m_\rho$, the gradient of the distance function is given by:

$$\nabla_x d_y = \frac{\rho^{-2} < x, y > x - y}{\sqrt{(\rho^{-1} < x, y >)^2 - \rho^2}} \tag{2}$$

HERACLES Method

Use automatic differentiation to simplify implementation

Use geoopt package that extends PyTorch to work with Riemannian manifolds



```
l = Logalike(rho=rho,
             character_matrix=cm,
             init_points=points,
             num_mutations=num_mutations,
             S=6,)

opt = geoopt.optim.RiemannianAdam(l.parameters(), lr=1e-3)
for i in range(num_cells):
    opt.zero_grad()
    loss = l.forward(Q, i)
    loss.backward()
    opt.step()
```

HERACLES Method

Currently finishing implementation

- Compute prior ancestral states π_a
- Update simulated data with Caissopea package
- Extract rate variables for infentiesmal generator Q

Acknowledgements

Massive thank you to Sitara and Prof Pe’er for guidance.

(And thanks to Anthony for his prior work and walking me through his code)